frame, said GB data packet flows having service requirements expressed in terms of bandwidth requirements, the duration of the first subframe being adjusted to satisfy the bandwidth requirements of said GB data packet flows, and (2) providing service to a second set of BE flows during a second subframe of the service frame, said BE data packet flows having service requirements that are not expressed in terms of guaranteed bandwidth requirements, the duration of the second subframe being the duration of the service frame minus the duration of the first subframe in the same service frame.

BRIEF DESCRIPTION OF THE DRAWINGS

In the drawings,

Fig. 1 shows an illustrative packet network including data sources, communication switches, and data destinations.

Fig. 2 shows an illustrative communication switch used in the packet network of Fig. 1.

Fig. 3A shows an example of pseudo-code used in a Deficit Round Robin (DRR) algorithm to handle flow *i* when a new packet arrives at the head of a flow queue.

Fig. 3B shows an example of pseudo-code used in a Surplus Round Robin (SRR) algorithm to update a timestamp *i* when the server completes the transmission of a packet.

Fig. 4 shows, in accordance with the present invention, a diagram illustrating the two-layered logical organization of the scheduler.

Fig. 5 shows a functional diagram of the queues, state tables, registers, and parameters utilized by the scheduler of the present invention.

Fig. 6 shows an illustrative block diagram of a particular implementation of the apparatus of Fig. 5.

C.H.
4/5/06

~~Figs. 7A and 7B~~ Figs. 7 and 7A-E show an illustrative flowchart describing a method of scheduling the transmission of packets in accordance with the present invention.

The global state table 507 stores data such as a global frame counter *GFC*, a reference timestamp increment T_Q^{PWS} for the PWS, and a reference timestamp increment T_Q^{SWS} for the SWS. The BE flow-aggregate state table 508 stores data that pertain to the BE flow aggregate, such as a timestamp F_{BE}, a BE running share ϕ_{BE}, and a BE cumulative share Φ_{BE}. A PWS First-In-First-Out (FIFO) queue 509 stores pointers to the GB flows 402. The PWS FIFO queue 509 indicates the order by which the PWS has to visit the GB flows to determine the transmission of their packets out of the respective GB flow queues 502. The registers 510 store pointers to the head and tail positions in the PWS FIFO queue 509. An SWS FIFO queue 511 stores pointers to the BE flows 405. The SWS FIFO queue 511 indicates the order by which the SWS has to visit the BE flows to determine the transmission of their packets out of the respective BE flow queues 505. The registers 512 store pointers to the head and tail positions in the SWS FIFO queue 511.

Figure 6 shows an illustrative block diagram of an input communication link interface 200 in which the scheduler may be utilized. The communication link interface 200 includes a data packet receiver 600, a scheduler 602, and a packet transmitter 601. Illustratively, the scheduler is shown to include a controller 603, a global state RAM 604, and registers 605, all on the same chip 606. A packet RAM 607 and a per-flow state RAM 608 are shown located on separate chips. Obviously, depending on the operating capacity and other characteristics, the scheduler 602 may be implemented in other configurations.

The controller 603 stores and runs the program that implements the method of the present invention. An illustrative example of the program that controls the operation of the communication link interface 200 is shown in flow-chart form in Figs. ~~7A-B.~~ With joint reference to Figs. 5 and 6, the packets in the

C. H.
4/6/06

7A-E

cumulative share Φ_{BE} is greater than zero. Every PWS service to the BE aggregate translates into an SWS service to a BE flow. Two events can trigger a reset of the BE cumulative share Φ_{BE} and therefore the end of the BE subframe: the last backlogged BE flow becomes idle, or the BE timestamp F_{BE} exceeds the PWS reference timestamp increment T_Q^{PWS}.

7A-E

C.H 4/5/06 Figures ~~7A and 7B~~ depict in flow-chart form a method for scheduling the transmission of packets according to the present invention. The flow chart is based on the assumption that SRR is the underlying scheduling algorithm. As far as functionality is concerned, there is no problem in using DRR instead of SRR. Similarly, the apparatus of Fig. 4 implements the PWS 401 and the SWS 404 using a single FIFO queue of backlogged flows for each of them (PWS FIFO queue 509 and SWS FIFO queue 511, respectively). Any other queueing structure that allows a clear separation of in-frame and out-of-frame flows could be used as well. Finally, the adoption of a WRR scheduler for handling the BE flows allows the enforcement of service fairness over the BE flows, but is not strictly required to achieve the efficient integration of GB and BE flows. Any other scheduling mechanism could be used as well to handle the BE flows in isolation from the GB flows.

7A-E

C.H. 4/5/06 The following description makes reference to Figs. 4, 5, 6, and ~~7A-B~~. The reference numbers to elements that are first defined in Fig. 4 (5, 6) begin with a 4 (5, 6), while the steps of the flow chart of Figs. ~~7A-B~~ are indicated by an S preceding the step number, e.g., S310. 7A-E

C.H. 4/5/06

In Fig. 7A, the scheduler 602 checks if there are newly received data packets in S310. If there are no newly received data packets in S310, and there are backlogged flows in S315, control passes to S500. If, instead, there are no newly received data packets in S310 and there are no backlogged flows in S315,